International Journal of Educational Research xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

International Journal of Educational Research

journal homepage: www.elsevier.com/locate/ijedures

Reducing the number of pupils in French first-grade classes: Is there evidence of contemporaneous and carryover effects?

Pascal Bressoux[a,*], Laurent Lima[a], Christian Monseur[b]

[a] Univ. Grenoble Alpes, Laboratory of Research on Acquisition in Context (LaRAC), France
[b] University of Liège, Belgium

ARTICLE INFO

Keywords:
Class size
Pupils' acquisitions
Growth curve modelling
Carryover effects

ABSTRACT

This study aims at investigating the contemporaneous and carryover effects of a class size reduction experiment. This experiment initially involved 100 reduced classes (average 10.5 pupils; range 8–12) that were contrasted with 100 full-size classes (average 21.3 pupils; range 15–27). Although the experiment was implemented only in Grade 1, the pupils have been followed up over a two-year period until the end of Grade 2. Multilevel growth curve modelling estimates show significant positive contemporaneous effect of the class size reduction in Grade 1, but no carryover effects during the following year. At the end of Grade 2, the effects have faded. Practical implications for a class-size reduction policy and limitations of the study are discussed.

1. Introduction

In 2017, the new French Government decided to implement a class size reduction policy in high priority educational areas (REP +): From the 2017–2018 school year, first-grade classes will have a maximum of 12 pupils. This policy will be extended to second-grade classes in the 2018–2019 school year and it should also be extended progressively to priority educational areas (REP).[1]

By so doing, the French Government launched an ambitious policy aimed at fighting against academic failure, which affects mainly pupils from disadvantaged background families. Indeed, France is one of the countries with the highest correlation between socio-economic status and academic achievement and one of the countries where the gap between high and low achievers is the widest (OECD, 2013a, 2013b). Of course, this policy is not only ambitious but also costly since according to the French Ministry, about 2500 first-grade classes will be divided by two. This implies that 2500 teachers will have to be sent in the newly created first-grade classes. The Ministry promised these 2500 teachers will not be novice teachers which means they will be redeployed from other schools or from different grades in the same school, and novice teachers will have to teach other grades.

It is therefore important to have an as clear as possible idea of what could be the outcomes of such a policy. In this paper we are able to report results from – and learn lessons from - an earlier French study, as yet unpublished, which addresses the effects of a similar class size reduction initiative. In 2002–2003, a class-size reduction experiment was conducted in France, which shares several similarities with the forthcoming policy: 1) the class-size reduction experiment was conducted in underprivileged areas (called "priority education zones"– ZEP – at that time), 2) in first-grade classes, and 3) the classes were reduced to 10 pupils on average (min

* Corresponding author at: Université Grenoble Alpes, UFR SHS, CS 40700, 38058 Grenoble Cedex 9, France.
E-mail address: pascal.bressoux@univ-grenoble-alpes.fr (P. Bressoux).

[1] Two types of priority education areas are identified: 1) The REP+ concerns districts or isolated sectors having the biggest concentrations of social difficulties with strong incidences on academic success and 2) the REP, more mixed socially, but meeting more social difficulties than those of the schools located outside the priority education. The REP+ labelling, and to a lower extent the REP labelling, provide more human and financial means to the schools.

https://doi.org/10.1016/j.ijer.2018.10.006
Received 6 August 2018; Received in revised form 11 October 2018; Accepted 15 October 2018
0883-0355/

8-max 12). In this article, we will analyze the data collected during this experiment to investigate two main questions: What are the effects of class size reduction? Does it have lasting effects? By examining pupils' growth curve literacy acquisition in reduced and full size classes, we will estimate the contemporaneous effects of class size reduction, that is, the effects during the class size reduction period (at first grade), and we will estimate the carry-over effects of class size reduction, that is, the effects during the following school year, once all the pupils were in full-size classes (at second grade).

2. Research on class size effects: is it worth reducing class size?

Research on class size has a long history (Glass & Smith, 1979; Slavin, 1989; Smith & Glass, 1980). Much has been written on the impact of class size on pupils' achievement and this has led to many controversies about the existence and magnitude of class size effects (Whitmore Schanzenbach, 2014). This debate is still alive (see, e.g., Blatchford et al., 2016; Li & Konstantopoulos, 2017; Sohn, 2016). Both experimental and naturalistic studies have been conducted in many countries (see, e.g., Ehrenberg, Brewer, Gamoran, & Willms, 2001), the two most representative studies of each kind being the STAR project in the USA (Finn & Achilles, 1999) and the CSPAR project in the UK (Blatchford, Bassett, Goldstein, & Martin, 2003). Both found negative effects of class size: pupils' achievement decreases as class size increases. Although not all the studies are conclusive about the effects of class size, overall, there is now general agreement about the existence of class size effects, although their size may vary amongst countries. In his mega-analysis, Hattie (2009) found an average effect-size $d = 0.21$. He argues that, compared to other potential factors, this effect size, although positive, is not large enough to promote class size reduction as a relevant policy to improve educational outcomes. According to him, this would only be one of those "politics of distraction" that appease parents, teachers and school leaders, but that are not worth the money spent (Hattie, 2015). Hattie reminds us that this is in line with the OECD Director for Education Andreas Schleicher's view: "If you have to choose between great teaching and a small class, go for the great teaching" (cited by Hattie, 2015, pp. 10–11).

According to Hattie, one of the reasons for this small effect is that teachers do not change their behavior when they move from large to small classes (Hattie, 2016). However, on the contrary, Finn and Shanahan (2016) argue that class size reduction has positive effects on its own: "Small classes in and of themselves produced benefits" (Finn & Shanahan, 2016, p. 133). This is because small classes provide teachers with more time of instruction (Finn & Shanahan, 2016, p. 133), pupils are more engaged in learning tasks (Finn, Panozzo, & Achilles, 2003) and individualized task-related interactions between teachers and pupils are more frequent (Blatchford, Bassett, & Brown, 2005).

These authors argue class-size benefits are not minor. First, global effects as calculated by meta-analysis techniques obscure the fact that the benefits vary according to different groups of students. In particular, they are much larger for low SES and minority students. Therefore, reducing class size might contribute to reduce inequalities (Finn & Shanahan, 2016). Second, it is not fair to compare class-size benefits to those provided by methods of teaching, because class size is not a process but a contextual factor: it is just a description of the number of pupils in a class (Blatchford, 2016). It does not say which teaching practices would be appropriate according to varying class sizes.

Unlike many other contextual factors (SES composition, free-meal percentage of pupils and so on), it is fairly easy to manipulate class size, assuming there is a political will – and the funding – to do it. In contrast, teaching practices are not that easy to manipulate at a large scale. Even though implementing efficacious practices, as revealed by randomized field trials, could in theory produce more gains than reducing class size, the experience shows that it is very difficult to implement these promising practices at a large scale (Bryk, 2014). Implementing evidence-based practices is very difficult and we still do not fully understand why. We still lack the knowledge of how to improve our educational systems and this is a big challenge for the 21st century (Bryk, 2014).

3. Class-size effects in France

Class size effects may vary between countries since every educational system has its own features which may affect the relation between class size and pupils' outcomes. Even though few studies have been specifically designed to study the effects of class size in France, some correlational studies provide results about the effects of class size on pupils' achievement. These results are not always congruent: some studies show non-significant effects; others show negative effects of class size (i.e., the bigger the class size the lower the academic performances); no studies show positive effects of large classes (Meuret, 2001). Monso (2014) recent review, based on the most robust studies controlling as best as possible for confounding factors, or following Angrist and Lavy (1999) instrumental variable approach (Bressoux, Kramarz, & Prost, 2009; Gary-Bobo & Mahjoub, 2013; Piketty & Valdenaire, 2006; Piketty, 2004), clearly concludes there is evidence that small classes have positive effects on pupil attainment in France.

It also appears that class size effects vary depending on the school level considered. Very few research has investigated the impact of class size at the pre-elementary level. To our best knowledge, Suchaut (1996) study is the only one. He found no effect of class size in kindergarten on precursor skills of future academic acquisitions such as memory, phonological awareness and space organization.

At elementary school, Mingat (1991) found a negative effect of large classes on reading acquisitions in grade 1 and a non-significant effect in mathematics. Leroy-Audouin and Mingat (1995) studied rural schools, most of which were multigrade. They found that, for third-grade pupils, classes with 18–22 pupils showed the largest progress, in comparison with classes that counted less than 18 pupils or more than 22 pupils. In a study that involved third to fifth grade classes, Bressoux (1994) found no significant effect of class size on pupils' progress in reading (grades 3–5) but he found a negative effect – only significant at a trend level ($p < .10$) – in multigrade classes. In another study that involved third grade, Bressoux (1996) found clear negative effects of class size on literacy and mathematics achievement.

Piketty (2004) showed that gains related to the reduction of class size at elementary school are larger for disadvantaged pupils. According to this author, reducing class size to 18 pupils in priority educational areas instead of the average of 21.9 at that time – with the consequence of increasing the size of classes in other areas to 24.16 pupils instead of an average of 23.3 – would reduce by 40% the performance gap between classes in priority educational areas and the other classes. He argued that such a decision would be highly equitable at no extra cost.

Bressoux et al. (2009) showed a clear and significant negative effect of class size on French and math achievement at grade 3. They also found evidence for heterogeneous effects of class size with a stronger effect for low achieving pupils in mathematics (but not in French), and also a stronger effect in the priority educational area schools in both French and math.

At the secondary level, Grisay (1993) found that, based on Sixth and Seventh-grade pupils' French and math achievement, the most effective schools were smaller schools and had smaller class sizes. However, another study found no significant effects of class size for French and math acquisitions at grade 6 (Schmitt Rolland & Thaurel-Richard, 1997; Verdon & Thaurel-Richard, 1997). Piketty and Valdenaire (2006) showed that, in middle schools, class size effects are approximately half the effect usually observed in elementary schools. They also found some evidence of negligible and non-significant effects in high schools.

Although pupils' achievement is the main focus of research on class size, Gary-Bobo and Mahjoub (2013) have studied the impact of class size on pupils' probability of promotion to the next grade. They found that the probability of promotion from grade 6 to grade 7 increases as class size diminishes. In contrast, in grades 8 and 9, the authors found no significant effects of class size on this probability.

Taken as a whole, these results tend to show negative effects of class size and these effects fade when grade increases. It is probably very important to acquire at early grades as good as possible basic competences for future learning. Therefore, it seems relevant to target a class size reduction policy to the first grades, that is, at the beginning of elementary school.

4. Class size long-term effects

Nowadays, one of the main points of debate is about long-term effects of class size (see, e.g. Whitmore Schanzenbach, 2016). These long-term effects may relate to cognitive test scores in various disciplines but they can also relate to later-life outcomes such as college completion, earnings, wages and even crime. Since the time of Project STAR (Finn, Gerber, Achilles, & Boyd-Zaharias, 2001) a growing body of research has emerged on this topic, whether reanalyzing the STAR data (e.g. Chetty et al., 2011; Dynarski, Hyman, & Whitmore Schanzenbach, 2013; Sohn, 2015) or analyzing new datasets (e.g., Fredriksson, Öckert, & Oosterbeek, 2013).

Project STAR is the unique large scale randomized class size reduction experiment, and much of our knowledge on class size effects rely on this study. Although it ended in Grade 3, achievement tests data were still collected (Finn, Gerber & Boyd-Zaharias, 2005). Several researchers who analyzed the STAR data claimed carryover effects because they discovered that the positive effects of class size reduction persisted beyond the experimental period (Sohn, 2015). At Grade 8, the size of the effect was overall one-third to one-half of the size of the effect that was observed during the experiment (Whitmore Schanzenbach, 2016). Finn et al. (2001) found that the longer the pupils were in small classes, the more benefits they had at Grade 8. Krueger and Whitmore (2001) analyzed student achievement in high school. They found that students assigned to small classes were more likely to take the college admissions tests (ACT or SAT) and scored higher. Krueger and Whitmore (2002) even found that black male students assigned to small classes were 0.6 points of percentage less likely to be convicted of a crime. However, Sohn (2015) has recently challenged the robustness of STAR cognitive carry over effects (grade 4 through to high school outcomes), arguing that most of these effects were produced by effective schools, which therefore confounded effects of class size reduction.

Chetty et al. (2011) tracked STAR pupils until age 27. They found that having attended small classes increased the probability to attend College. They failed to find a relationship with earnings at age 27 but found that African-American people had an earnings increase of $250 (Whitmore Schanzenbach, 2016).

Dynarski et al. (2013) followed STAR pupils through to approximately age 30. They found that assignment to a small class increased the probability of attending college by 2.7 percentage points, with much larger effects amongst Black and disadvantaged people. Furthermore, small class attendees were more likely to complete a college degree and to shift towards high-earning fields such as STEM (science, technology, engineering and mathematics), business and economics.

In the UK, Iacovou (2002), using longitudinal data from the National Child Development Study, found that the effects of class size observed at age 7, namely higher attainment in reading in smaller classes, persisted through to age 11 not for all groups of pupils but for some of them: girls and children from larger families.

In Sweden, Fredriksson et al. (2013) found that people who attended smaller classes in the last three years of elementary school (age 10–13) perform better on cognitive tests at age 13, 16 and 18, are more likely to have completed at least a Bachelor's degree and have an increase in wages when they are aged 27–42.

Leuven and Løkken (2017) did not find these effects on Norwegian data. They found no significant impact of class size in primary school on earnings and completed schooling. The results obtained in middle school are in line with those for primary school.

To our best knowledge, no studies in France have analyzed the long-term effects of class size. So we had no information on this point before the present study.

5. The present study

Our study aims to estimate the contemporaneous and carryover effects of a class-size reduction experiment in France. As described above, the French Ministry of Education instigated in 2002–2003 a class-size reduction experiment which involved overall

200 First-grade classes: 100 experimental classes which size was dramatically reduced (average 10.5 pupils; range 8–12) were contrasted with 100 full-size classes (average 21.3 pupils; range 15–27) used as control. All schools were located in priority educational areas in different regions in France.

It is important to note that teachers and pupils were not assigned to one group or another under a strict random process. The Ministry gave instructions for reaching equilibrated samples but some initiatives were left to local authorities in the assignment process.

After the experiment was finished, all students were assigned to full-size second-grade classes. The students were followed up from the beginning of Grade 1 until the end of Grade 2.

Even though the time span of our study does not allow us to estimate the longer term effects of human capital formation, it is one of the rare studies that 1) brings evidence on a dramatic reduction of class size (much larger than the STAR reduction), and 2) estimates contemporaneous and carryover effects of class size reduction. Few studies worldwide have such features, and to date none in France.

6. Method

6.1. Participants

The study initially comprised 100 reduced-size classes and 100 full-size classes distributed in ten academic regions all over France (Aix-Marseille, Amiens, Créteil, Lille, Lyon, Paris, La Réunion, Rouen, Strasbourg, Versailles). All these classes belong to priority educational areas.[2] The experimental group and the control group each initially consisted of about 1000 students. The equal number of pupils between the two groups, despite classes twice larger in the control group, is due to the fact only 10 randomly chosen students were assessed in the control classes.

At the first assessment (Grade 1, September), 2047 pupils were tested, 1761 at the second assessment (Grade 1, February), 1566 at the third assessment (Grade 1, May), 1481 at the fourth assessment (Grade 2, September) and 1266 at the fifth assessment (Grade 2, May). In longitudinal studies, missing data are unavoidable, especially when pupils are followed during two years and repeatedly tested. However, in this study, attrition is greater than expected due to a particular social context in France. At spring of the 2002–2003 school year, that is at the time of the last Grade 1 assessment period, many teachers went on strike. So did some of the participant teachers who, as a consequence, refused to participate in the last assessment phase.

Since we do not know exactly the reason why teachers decided to go or not to go on strike, nor do we know to what extent the progress of the class was disrupted by these events, we deemed it was unrealistic to assume the missing data are random, which is an assumption to impute data for these pupils. Since these classes were excluded from the Grade 2 follow-up, we decided to retain in the analyses all the pupils who were present at least at the first (beginning of Grade 1) and at the last (end of Grade 2) assessment phases. In this sample, not all the pupils have been tested at every assessment phase, but we can assume that missing data are due to a random process. Indeed, the pupils have been part of the study during the whole period and the missing data depend on whether they were present or absent during the multiple assessment phases. For those pupils who have missing data between the first and last testing phases, the multilevel growth curve model imputes data assuming a missing at random process. On average, pupils have been tested more than four times each.

In total, our analyses were carried out on 1264 pupils, 663 of which belong to the experimental group and 601 of which belong to the control group. They are located in 69 schools and 149 Grade 1 classes (81 experimental classes and 68 control classes). This corresponds to a total of 5822 observations (number of measures).

6.2. Materials and procedure

Over the two-year period, the pupils' literacy achievement was assessed 5 times (at the beginning, middle and end of grade 1; at the beginning and end of grade 2). For each phase, tests were constructed by a team of cognitive psychologists in order to assess pupils' literacy competence. Five skills were assessed: word recognition, vocabulary, spelling, phonology, reading comprehension (Gombert, 2003). Tests were administered collectively in the class under the teacher supervision, and were marked by the Ministry of Education. A global literacy competence score was calculated at each phase. The internal consistencies of the tests are respectively $\alpha = 0.89$ at T1, $\alpha = 0.96$ at T2, $\alpha = 0.97$ at T3, $\alpha = 0.97$ at T4 and $\alpha = 0.94$ at T5.

Characteristics of pupils (SES measured through father's occupation INSEE classification 2003; gender) and teachers (job; post and Grade 1 teaching experiences) were collected through questionnaires.

6.3. Methodological considerations

Anchor items were included in the tests administered at the different time periods. The cognitive data were scaled with Conquest software (Wu, Adams, & Wilson, 1997) according to a one-parameter logistic Item Response Theory model. The pupil proficiency estimates (Weighted Maximum Likelihood Estimate or *Warm* estimate) for each assessment phase are therefore on a single scale. The progress in literacy skills can consequently be estimated.

[2] At that time, there was no distinction between REP+ and REP.

Table 1
Descriptive statistics of the pupils.

	Control group		Experimental group	
	N	Percentage	N	Percentage
Craftsman	25	4.16	27	4.07
Manager or senior executive	16	2.66	8	1.21
Intermediate profession	52	8.65	41	6.18
Office worker	74	12.31	59	8.90
Worker	201	33.44	290	43.74
Other	206	34.28	222	33.48
Boy	313	52.08	348	52.57
Girl	288	47.92	314	47.43

Multilevel growth curve models were performed, using SAS software, to estimate the shape and rate of pupils' achievement growth across the two years depending on their assignment to the experimental or control group. Time is measured in number of months (0, 5, 8, 12, 20) so the intercept represents the achievement level at the beginning of the experiment.

We specified multilevel growth curve models as piecewise regression in order to assess the acquisition growth rate of the two groups in function of different periods: during first grade (i.e. from time 0 to time 8), during summer vacations (i.e. from time 8 to time 12) and during second grade (i.e. from time 12 to time 20). This will allow us to estimate different slopes for each group and each period.

7. Results

7.1. Descriptive statistics

In terms of pupils' characteristics, the two groups are close to each other, except that there are slightly more workers in the experimental group (see Table 1).

In terms of literacy scores, the two groups are very close to each other (see Table 2). The control group has a slightly higher initial score but this may be due to the fact that there are more workers in the experimental group, and this category is known for having relatively low achievement scores. At the third assessment, the experimental group seems to have a higher score compared to the control group, and both groups have equivalent scores at the final assessment.

As for the teachers' characteristics, the two groups are fairly different because experimental teachers are less experienced than those in the control group (Table 3). This is true for all three indicators although the most striking differences are for the number of years of teaching experience and the number of years of first-grade teaching experience. Obviously, we will need to control for these differences in our models in order to have non-biased class size estimates.

Table 2
Descriptive statistics of the literacy scores.

Variable	N	Mean	Standard deviation	Minimum	Maximum
Control group					
Score 1	601	−0.67	1.15	−3.67	3.55
Score 2	530	1.00	0.97	−1.24	5.85
Score 3	510	1.74	1.11	−0.80	5.76
Score 4	498	1.92	1.18	−0.97	5.76
Score 5	601	2.88	1.13	−0.33	6.39
Experimental group					
Score 1	663	−0.77	1.16	−3.39	3.55
Score 2	588	1.06	0.93	−1.45	4.72
Score 3	572	1.83	1.10	−2.50	4.76
Score 4	596	1.98	1.17	−1.16	5.17
Score 5	663	2.87	1.13	−0.19	6.39

Table 3
Descriptive statistics of the teachers.

Variable	N	Mean	Standard deviation	Minimum	Maximum
Control group					
Teaching experience	68	13.72	11.69	0	39.00
Length of time in post	66	6.74	8.18	0	37.00
First-grade teaching experience	68	7.26	8.66	0	37.00
Experimental group					
Teaching experience	81	8.96	8.36	0	35.00
Length of time in post	80	4.29	5.89	0	27.00
First-grade teaching experience	81	3.63	4.79	0	24.00

7.2. Growth curve modelling

We specified a three-level model with literacy scores at level one, pupils at level 2 and schools at level 3. Since we do not have information on Grade 2 class assignment, we could not include classes as a specific level of analysis.

Tables A1 and A2 in Appendix A describe the model specification process. We can see how the piecewise models (Table A2, models 3 and 4) improve the fit to the data compared to the linear unconditional growth curve model (Table A1, model 2). While the linear unconditional growth curve model estimates a 7% between school variance, a 50% between pupil variance and a 43% between measure variance, the piecewise unconditional growth curve model gives a 7%, 62%, 31% distribution. The between-measure variance is lower than estimated in the linear model because progress is not linear. Model 4 gives a better fit because it includes time variance and intercept/time covariance at both levels 2 and 3, indicating that the rate of change varies significantly between pupils and also between schools. The rate of change is negatively correlated with initial achievement level.

Model 5 (Table 4) is the final model including all our independent variables. The random part of the model shows that, even after controlling for pupil, class and teacher characteristics, there remains residual variation amongst pupils in both initial achievement level and rate of growth. The same result appears at the school level: not all the schools have the same achievement level and the rate

Table 4
Parameter estimates and their standard errors of the final piecewise growth curve model.

Parameters	Model 5
Fixed Effects	
Intercept	−0.361 (0.141)*
Time	0.297 (0.007)***
(Time – 8)*post-Grade 1	−0.243 (0.014)***
(Time – 12)*Grade 2	0.067 (0.014)***
Fathers' occupation (reference = manager or senior executive)	
Craftsman	−0.399 (0.175)*
Intermediate profession	−0.1575 (0.151)
Office worker	−0.319 (0.143)*
Worker	−0.556 (0.127)***
Other	−0.602 (0.128)***
Girl	0.240 (0.053)***
Reduced class	0.023 (0.097) ns
Grade 1 experience	0.006 (0.006) ns
Time*Reduced class	0.0295 (0.0075)***
(Time – 8)*Post-grade 1*Reduced class	−0.0641 (0.0154)***
(Time – 12)*Grade 2*Reduced class	0.0225 (0.0158) ns
Time*Grade 1 experience	0.0019 (0.0005)***
(Time – 8)*Post-grade 1*Grade 1 experience	−0.0050 (0.0011)***
(Time – 12)*Grade 2*Grade 1 experience	0.0034 (0.0011)**
Random effects	
Between-school variance (level 3)	
Intercept variance	0.140 (0.037)***
Intercept/Time covariance	−0.0054 (0.0017)**
Time variance	0.0005 (0.0001)***
Between-student variance (level 2)	
Intercept variance	0.751 (0.039)***
Intercept/Time covariance	−0.0065 (0.0015)***
Time variance	0.0011 (0.0001)***
Between-measure variance (level 1)	0.277 (0.007)***
–2 log L (deviance)	12,353.10

* $p < .05$.
** $p < .01$.
*** $p < .001$.

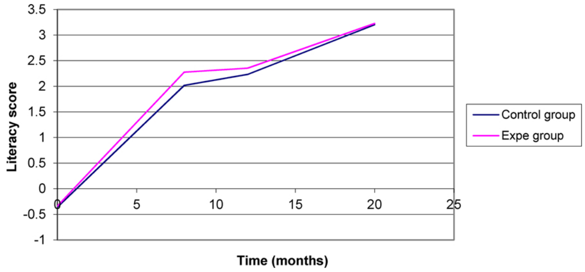

Fig. 1. Literacy score growth curves of the experimental and control groups.

of growth varies significantly between schools. The negative covariance between the intercepts and the slopes indicates that progress is higher in low-achieving schools. We must remember, however, that all the schools are in priority educational areas and there are no high performing schools in our sample. It might also be that the result just reflects a classical regression to the mean effect. Indeed, a negative covariance between intercepts and slopes was also found at the pupil level.

We present now the fixed effect part of the model, which includes class size. Pupils whose father's occupation is manager or senior executive are higher achievers on average than the other pupils. The gap is particularly marked with workers. Girls also tend to perform better than boys.

Having been at grade 1 with a more experienced teacher (see Time*Grade 1 experience coefficient) improves the growth rate in literacy scores during grade 1. This effect disappears at the end of Grade 2.

As for the effects of class size reduction, the results of the piecewise growth curve model are presented in Fig. 1, for ease of interpretation. We can see that both groups have an equivalent average achievement at the beginning of the experiment, but pupils progress differently according to the group they belong to. During the eight first months, the rate of change is higher in the experimental group compared to the control one: there is an increasing gap in literacy scores in favor of the experimental group. At month 8, the gap is the largest and the effect size is $d = 0.22$. There is evidence of contemporaneous effects of the class size reduction: small class attendees progress more in literacy skills than their regular-class counterparts at Grade 1.

During the summer vacation (months 8–12), the growth rate diminishes for both groups. We note also that the gap between the two groups is lower at the beginning of Grade 2. During Grade 2, the gap continues to diminish and, by the end of Grade 2, the effect of reduced-size classes has faded away. Therefore, there is no evidence of carryover effects.

7.3. Are our results affected by a Hawthorne effect?

Sohn (2015) wondered if there was a Hawthorne effect that might affect Project STAR results. The Hawthorne effect is the process where individuals in an experimental group increase their productivity simply because they are being studied. By nature, the design of a class size reduction experiment does not place reduced-class teachers in a blind test situation: they do know they are in the experimental group. This might foster their will to improve pupils' learning, resulting in higher achievement, not because of the class size reduction itself but because of teachers' higher motivation and commitment. Sohn's estimations lead him to conclude that a Hawthorne effect is not a convincing explanation of the STAR results.

If the advantage of the experimental group over the control one was due to a Hawthorne effect rather than to class size, then class size would have no effect, or minor effects in the control group. In our study, the class-size variation in the control group (from 15 to 27) is large enough so we can estimate its effect on the literacy scores at the end of Grade 1. We tested this hypothesis in an analysis restricted to the control group. The model we used is equivalent to model 5 (see Table 4) except that the dummy variable "reduced class" was replaced by the variable "class size" (i.e. number of pupils per class).

The estimated coefficient of class size in the control group equals -0.031, significant only at a trend level ($p < .09$). However, due to the small sample, the effect size matters more here than significance. In the control group, a one-pupil reduction in a class entails gains of 3.1 percentage point of a standard deviation. We then compared this coefficient with the one obtained using the same model estimated on the whole sample (i.e. including the experimental and the control groups). This coefficient equals -0.022 and it is significant ($p < .05$). Since the two coefficients are relatively close to each other, the coefficient estimated on the control group being even slightly higher than the coefficient estimated on the whole sample, there is no reason for suspecting any Hawthorne effect in the results produced by the class-size reduction experiment.

8. Discussion

This article aimed at evaluating the effects of a class size reduction experiment. By using a piecewise growth curve model approach, we were able to estimate the impact of reducing class size from an average of 21 to an average of 10 on the rate of change in literacy scores over two years including Grade 1, the experiment period, and Grade 2, the year pupils returned to regular classes. This allowed us to investigate both contemporaneous and carryover effects of the experiment.

Our results clearly show that pupils who were assigned to small classes in Grade 1 made greater progress during this year; their rate of change was significantly higher compared to that of pupils who attended regular classes. By the end of Grade 1, small-class pupils' literacy score was 22 percentage points of a standard deviation higher than the control pupils' score. This effect, which corresponds to a 2 percentage points increase for one child less in a class (because there is an 11 pupil difference on average between the experimental and the control groups), is approximately the same as the one found on French scores at Grade 2 by Piketty and Valdenaire (2006), who used an instrumental variable approach.

In contrast, we could not detect any significant carryover effects. By the end of Grade 2, the advantage small class attendees had at Grade 1 was lost. To our best knowledge, this study is the first to explore lasting effects of class size in France. In appearance, our results seem contradictory with those from Project STAR which stated there were lasting benefits of class size reduction (Finn et al., 2001). However, these results were obtained for pupils who spent several years (up to four years) in small-size classes, but not for those who spent only one year, as in our study. If one restricts the comparison to STAR pupils who attended reduced classes only one year, our results are actually consistent with those of Project STAR: "One year (in kindergarten or grade 1) produces early gains but does not produce lasting effects" (Finn et al., 2001, p. 174). Our results are also consistent with those from Blatchford et al. (2003). In their study, they showed that the gains made in the small size reception classes were still observable by the end of year 1, but greatly eroded in math, and dependent on whether pupils moved to smaller, similar or larger year 1 classes in literacy. By the end of year 2, the gains were lost in both subjects. It seems that, to observe lasting effects it is necessary that pupils spend more than one year in a small-size class. According to Ramey and Ramey (1998), one of the mechanisms that is susceptible to maintain a higher rate of acquisition is when "enhanced knowledge results in greater environmental opportunities provided by others" (Ramey & Ramey, 1998, p. 119). This may not be the case for first-grade small-class attendees when they enter Grade 2. Indeed, the second-grade teachers may not provide greater learning opportunities to reduced class attendees because they need to adapt their teaching to all their pupils, some of which may come from first-grade reduced classes and others from first-grade regular-classes. Furthermore, it is plausible that only one year in a reduced class does not produce an increase in pupils' intellectual skills that is sufficient to gain more from later learning opportunities in less favorable environments (i.e. non-reduced classes).

9. Practical considerations

As for the French Government policy of class size reduction, our results demonstrate that reducing class size does matter and can improve pupils' literacy skills; at least, it can produce early gains. Our results also demonstrate that, in terms of cognitive abilities, one year in a small class is not enough to reveal enduring effects. Therefore, the ambition of the French Government to extend the class size reduction to Second-grade classes is probably necessary.

The model estimates also show that the positive contemporaneous effect of the class size reduction can be totally erased if reduced classes are taught by novice teachers. Therefore, we strongly recommend to act so that the mechanical increase of First-grade teachers is not accompanied by an overall reduction of teachers' experience in reduced classes.

It should be noted that, in our study, experimental classes are much smaller than in most previous studies, and even the control classes are not that large compared to non-European countries. For example, average class size in most eastern countries is much higher (Blatchford et al., 2016; Galton, Lai, & Chan, 2015), so some of our results and considerations may not apply directly to these contexts. Nonetheless, our study shows that even in a context where class size is limited, its reduction still has beneficial effects.

10. Limits of our study

In terms of causal reasoning, a major limitation of our study is that it is not a randomized experiment. One can never be sure that there is no hidden factor that hinders the results. However, both our contemporaneous effects and carryover effects are congruent with the STAR results, which was a random experiment. Furthermore, our contemporaneous effects are congruent in terms of direction and magnitude with those obtained in France by Piketty (2004) using an instrumental variable approach, by so doing considerably limiting the risks of confounding factors or endogeneity problems. This makes us believe our results are highly plausible.

We do not have information on the Second-grade class sizes, which forces us to assume that there are no differences in class size between the two groups so this does not interfere with our carryover results. Although this assumption seems very likely, it prevents us from investigating the impact of having been in a more or less big class during year 2 conditional to the group belonging during year 1. Obviously, our results would have been richer if we had this information.

The carryover effects of the experiment are evaluated just one year after the experiment ended. This can be considered a relatively short period and we have no information on subsequent academic success nor on other life outcomes. Whitmore Schanzenbach (2016) underlines that, in the STAR results, long-term impacts are larger than what would have been predicted on the short-run test score gains. She argues that this suggests some benefits of small classes are not directly captured by cognitive tests.

Finally, our study does not investigate the costs of the class size reduction, and whether or not other policies may be more cost-effective.

Declarations of interest

None.

Acknowledgement

This work was funded by the Ministère de l'Education Nationale – Direction de la Programmation et du Développement.

Appendix A

Table A1
Empty model and unconditional linear multilevel growth curve model.

Parameters	Model 1	Model 2
Fixed Effects		
Intercept	0,379 (0,047)***	−0,147 (0,049)**
Time		0,170 (0,002)***
(Time – 8)*post-Grade 1		
(Time – 12)*Grade 2		
Random effects		
Between-school variance (level 3)		
Intercept variance	0,090 (0,027)***	0,094 (0,027)***
Between-student variance (level 2)		
Intercept variance	0,335 (0,037)***	0,703 (0,035)***
Between-measure variance (level 1)	2,370 (0,051)***	0,610 (0,013)***
–2 log L (deviance)	20760,97	14951,46

*p < .05.
** p < .01.
*** p < .001.

Table A2
Piecewise multilevel growth curve models.

Parameters	Model 3	Model 4
Fixed Effects		
Intercept	−0,684 (0,049)***	-0,683 (0,056)***
Time	0,323 (0,003)***	0,323 (0,004)***
(Time – 8)*post-Grade 1	−0,304 (0,009)***	−0,305 (0,007)***
(Time – 12)*Grade 2	0,097 (0,009)***	0,098 (0,008)***
Random effects		
Between-school variance (level 3)		
Intercept variance	0,090 (0,026)***	0,144 (0,037)***
Intercept/Time covariance		−0,0052 (0,0017)**
Time variance		0,0005 (0,0001)***
Between-student variance (level 2)		
Intercepts variance	0,749 (0,036)***	0,791 (0,041)***
Intercept/Time covariance		−0,0062 (0,0016)***
Time variance		0,0011 (0,0001)***
Between-measure variance (level 1)	0,379 (0,013)***	0,282 (0,007)***
–2 log L (deviance)	12907,32	12468,15

*p < .05.
** p < .01.
*** p < .001.

References

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quaterly Journal of Economics, 114*(2), 533–574.
Blatchford, P. (2016). Is it true that class size does not matter. A critical review of research on class size effects. In P. Blatchford, K. W. Chan, M. Galton, K.-C. Lai, & J. Chi-Kin Lee (Eds.). *Class size: Eastern and Western perspectives* (pp. 92–104). London-New York: Routledge.
Blatchford, P., Bassett, P., & Brown, P. (2005). Teachers' and pupils' behavior in large and small classes: A systematic observation study of pupils aged 10 and 11 years. *Journal of Educational Psychology, 97*(3), 454–467.
Blatchford, P., Bassett, P., Goldstein, H., & Martin, C. (2003). Are class size differences related to pupils' educational progress and classroom processes? Findings from the Institute of Education class size study of children aged 5-7 years. *British Educational Research Journal, 29*(5), 709–730.
Blatchford, P., Chan, K. W., Galton, M., Lai, K.-C., & Chi-Kin Lee, J. (Eds.). (2016). *Class size: Eastern and Western perspectives*. London-New York: Routledge.
Bressoux, P., Kramarz, F., & Prost, C. (2009). Teachers' training, class size and students' outcomes: Learning from administrative forecasting mistakes. *The Economic Journal, 119*, 540–561.
Bryk, A. S. (2015). 2014 AERA distinguished lecture. Accelerating how we learn to improve. *Educational Researcher, 44*(9), 467–477.
Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics, 126*, 1593–1660.
Dynarski, S., Hyman, J., & Whitmore Schanzenbach, D. (2013). Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion. *Journal of Policy Analysis and Management, 32*(4), 692–717.
Ehrenberg, R. G., Brewer, D. J., Gamoran, A., & Willms, J. D. (2001). Class size and student achievement. *Psychological Science in the Public Interest, 2*(1), 1–30.

Finn, J. D., & Achilles, C. M. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis, 21*(2), 97–109.
Finn, J. D., Gerber, S. B., Achilles, C. M., & Boyd-Zaharias, J. (2001). The enduring effects of small classes. *Teachers College Record, 103*(2), 145–183.
Finn, J. D., Gerber, S. B., & Boyd-Zaharias, J. (2005). Small Classes in the Early Grades, Academic Achievement, and Graduating From High School. *Journal of Educational Psychology, 97*(2), 214–223.
Finn, J. D., Panozzo, G. M., & Achilles, C. M. (2003). The "why's" of class size: Student behavior in small classes. *Review of Educational Research, 73*(3), 321–368.
Finn, J. D., & Shanahan, M. E. (2016). Does class size (still) matter? In P. Blatchford, K. W. Chan, M. Galton, K.-C. Lai, & J. Chi-Kin Lee (Eds.). *Class size: Eastern and Western perspectives* (pp. 121–145). London-New York: Routledge.
Fredriksson, P., Öckert, B., & Oosterbeek, H. (2013). Long-term effects of class size. *The Quarterly Journal of Economics, 128*(1), 249–285.
Galton, M., Lai, K. C., & Chan, K. W. (2015). *Learning to teach small classes: Lessons from East Asia.* Oxon: Routledge.
Gary-Bobo, R. J., & Mahjoub, M. B. (2013). Estimation of class-size effects, using "Maimonides' rule" and other instruments: The case of French junior high schools. *Annals of Economics and Statistics,* 111–112.
Glass, G. V., & Smith, M. L. (1979). Meta-analysis of research of class size and achievement. *Educational Evaluation and Policy Analysis, 1*(1), 2–16.
Gombert, J.-E. (2003). *Expérimentation CP à effectif réduit : évaluation des élèves, traitements phonologiques, morphologiques et orthographiques. Rapport final de la convention n° 2003-003*.
Grisay, A. (1993). Le fonctionnement des collèges et ses effets sur les élèves de sixième et de cinquième. *Dossiers Education et Formations, 32*.
Hattie, J. (2009). *Visible learning.* London, New York: Routledge.
Hattie, J. (2015). *What doesn't work in education: The politics of distraction. Open ideas at pearson.* ISBN: 9780992422677.
Hattie, J. (2016). The right question in the debates about class size. Why (positive) effect so small? In P. Blatchford, K. W. Chan, M. Galton, K.-C. Lai, & J. Chi-Kin Lee (Eds.). *Class size: Eastern and Western perspectives* (pp. 105–118). London-New York: Routledge.
Iacovou, M. (2002). Class size in the early years: Is smaller really better? *Education Economics, 10*(3), 261–290.
INSEE (2003). *Nomenclature des professions et catégories sociales.* https://www.insee.fr/fr/metadonnees/pcs2003/categorieSocioprofessionnelleAgregee/1?champRecherche=true.
Krueger, A., & Whitmore, D. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. *The Economic Journal, 111*, 1–28.
Krueger, A. B., & Whitmore, D. (2002). Would smaller classes help close the black-white achievement gap? In J. Chubb, & T. Loveless (Eds.). *Bridging the Achievement Gap* (pp. 11–46). Washington, DC: Brookings Institution Press.
Leroy-Audouin, C., & Mingat, A. (1995). *L'école primaire rurale en France. Rapport pour la Direction de la Prévision du Ministère des Finances*.
Leuven, E., & Løkken, S. (2017). *Long term impacts of class size in compulsory school. IZA discussion paper n° 10594*.
Li, W., & Konstantopoulos, S. (2017). Does class-size reduction close the achievement gap? Evidence from TIMMS 2011. *School Effectiveness and School Improvement, 28*(2), 292–313.
Meuret, D. (2001). *Les recherches sur la réduction de la taille des classes. Rapport pour le HCEE*.
Mingat, A. (1991). *Expliquer les acquisitions au cours préparatoire : les rôles de l'enfant, de la famille et de l'école. Revue Française de Pédagogie, vol. 95*, 47–63.
Monso, O. (2014). L'effet d'une réduction de la taille des classes sur la réussite scolaire en France : développements récents. *Education et Formations, 85*, 47–61.
OECD (2013a). *PISA 2012 results: What makes schools successful? resources, policies and practices (Volume IV).* PISA: OECD Publishinghttps://doi.org/10.1787/9789264201156-en.
OECD (2013b). *PISA 2012 results: Excellence through equity: Giving every student the chance to succeed (Volume II).* PISA: OECD Publishing.
Piketty, T. (2004). *L'impact de la taille des classes et de la ségrégation sociale sur la réussite scolaire dans les écoles françaises. Une estimation à partir du panel primaire 1997.* Paris School of Economics, mimeo.
Piketty, T., & Valdenaire, M. (2006). L'impact de la taille des classes sur la réussite scolaire dans les écoles, collèges et lycées français. *Les Dossiers Evaluations et Statistique, 173*.
Ramey, C. T., & Ramey, S. (1998). Early intervention and early experience. *The American Psychologist, 53*, 109–120.
Schmitt Rolland, S., & Thaurel-Richard, M. (1997). Pratiques pédagogiques de l'enseignement du français en sixième et progrès des élèves. *Dossier Education et Formations, 87*.
Slavin, R. (1989). Class size and student achievement: Small effects of small classes. *Educational Psychologist, 24*(1), 99–110.
Smith, M. L., & Glass, G. V. (1980). Meta-analysis of research on class size and its relationship to attitudes and instruction. *American Educational Research Journal, 17*(4), 419–433.
Sohn, K. (2015). Nonrobustness of the carryover effects of small classes in Project STAR. *Teachers College Record, 117*, 1–26.
Sohn, K. (2016). A review of research on Project STAR and path ahead. *School Effectiveness and School Improvement, 27*(2), 116–134.
Suchaut, B. (1996). *Le temps scolaire : Allocation et effets sur les acquisitions des élèves en Grande section et au CP. Unpublished doctoral dissertation.* University of Burgundy.
Verdon, R., & Thaurel-Richard, M. (1997). Pratiques pédagogiques de l'enseignement des mathématiques en sixième et progrès des élèves. *Dossier Education et Formations, 84*.
Whitmore Schanzenbach, D. (2014). *Does class size matter?* Retrieved [March 30, 2015] fromBoulder, CO: National Education Ploicy Center. http://nepc.colorado.edu/publication/does-class-size-matter.
Whitmore Schanzenbach, D. (2016). Long-term impacts of class size reduction. In P. Blatchford, K. W. Chan, M. Galton, K.-C. Lai, & J. Chi-Kin Lee (Eds.). *Class size: Eastern and Western perspectives.* (pp. 59–79). London-New York: Routledge.
Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ConQuest: Multi-Aspect Test Software [Computer software].* Camberwell: Australian Council for Educational Research.